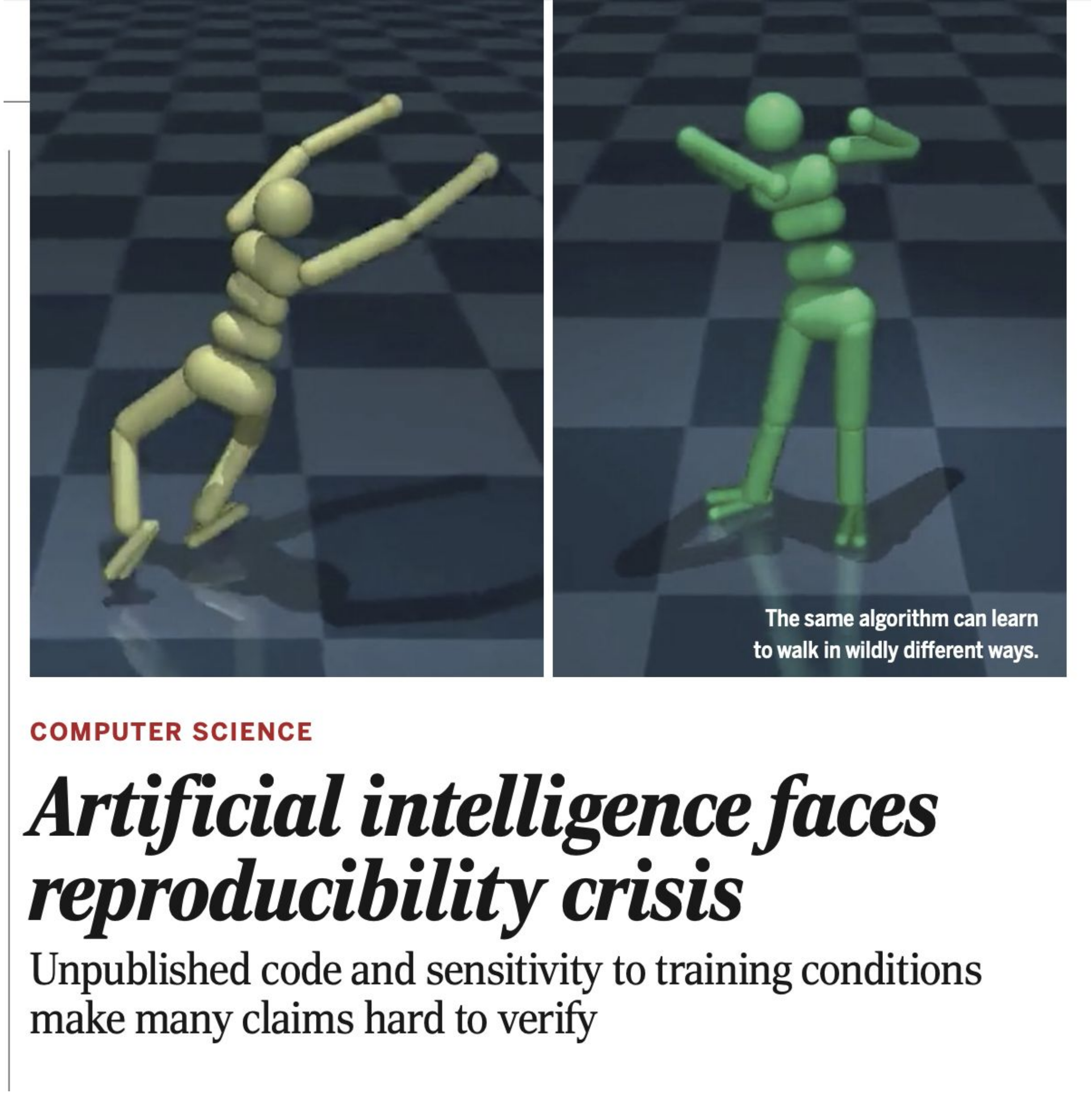
Reproducibility in AI, and What Computing Professionals Should Know for Supporting Researchers

SDSC SAN DIEGO SUPERCOMPUTER CENTER

UC San Diego



The same algorithm can learn to walk in wildly different ways.

COMPUTER SCIENCE

Artificial intelligence faces reproducibility crisis

Unpublished code and sensitivity to training conditions make many claims hard to verify

Hutson, Matthew. “Artificial Intelligence Faces Reproducibility Crisis.” *Science*, vol. 359, no. 6377, Feb. 2018, pp. 725–26. *DOI.org*

What is Reproducibility?

- **Confusion between:**
  - Repeatability
  - Replicability
  - Reproducibility

- **Definitions can differ between scientific disciplines.**

- **Definitions can change over time as the literature evolves.**

Analytical Chemistry

- Within-run: Same staff, same lab, same solutions, same equipment, same procedure and within a short period of time (same lab conditions).

- Between-run: Different staff, different lab, different equipment, same procedure (or as close as possible) and at a different time.

- **Repeatability** describes the precision of within-run replicates.
- **Reproducibility** describes the precision of between-run replicates.

- The reproducibility of a method is normally expected to be poorer (i.e. with larger random errors) than its repeatability.

Miller, J. N., and J. C. Miller. *Statistics and Chemometrics for Analytical Chemistry*. 6. ed, Prentice Hall, 2010.

ACM 1.0 - June 2016

- Repeatability (Same team, same experimental setup) - *Repeatability in Analytical Chemistry*
- Reproducibility (Different team, different experimental setup) - *Not defined in AC*
- Replicability (Different team, same experimental setup) - *Reproducibility in Analytical Chemistry*

Inspired by the International Vocabulary for Metrology (VIM).

https://www.acm.org/publications/policies/artifact-review-badging

ACM 1.1 - August 2020

- Repeatability (Same team, same experimental setup) - *Repeatability in Analytical Chemistry*

- Reproducibility (Different team, same experimental setup) - *Reproducibility in AC*

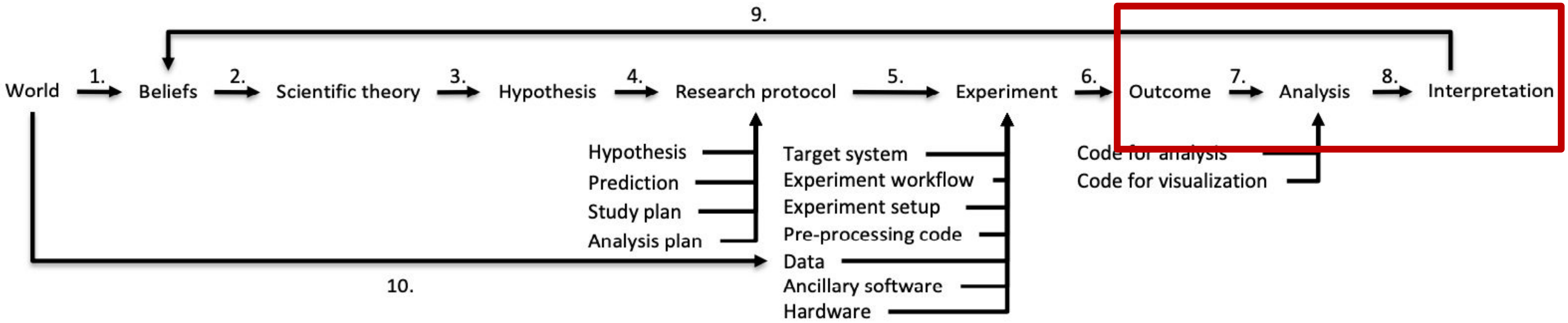
- Replicability (Different team, different experimental setup) - Not defined in Analytical Chemistry

As a result of discussions with the National Information Standards Organization (NISO), it was recommended that ACM harmonize its terminology and definitions with those used in the broader scientific research community.

https://www.acm.org/publications/policies/artifact-review-and-badging-current

The Scientific Method

Perhaps a better way to think of reproducibility to examine the scientific method:



The scientific method as a ten step process: 1) observe the world to form beliefs about it; 2) explain causes and effects by forming a scientific theory; 3) formulate a genuine test of the theory; 4) design an experiment to test the theory; 5) implement the experiment; 6) conduct the experiment; 7) analyse the outcome; 8) interpret the analysis; 9) update beliefs according to the result; and 10) observe the world systematically.

Gundersen, Odd Erik. “The Fundamental Principles of Reproducibility.” *ArXiv:2011.10098 [Cs]*, Nov. 2020. *arXiv.org*,

Reproducibility Based on the Scientific Method

- **Outcome reproducible**: The outcome of the reproducibility experiment is the same as the outcome produced by the original experiment. When the outcome is the same, the same analysis and interpretation can be made, which leads to the same result and hence the hypothesis is supported by both experiments. The experiment is outcome reproducible.
- **Analysis reproducible**: The outcome of the reproducibility experiment does not have to be the same as the outcome produced by the original experiment, but as long as the same analysis can be made and it leads to the same interpretation, the experiment is analysis reproducible.
- **Interpretation reproducible**: Neither the outcome nor the analysis need to be the same as long as the interpretation of the analysis leads to the same conclusion. In this case the experiment is interpretation reproducible.

Gundersen, Odd Erik. “The Fundamental Principles of Reproducibility.” *ArXiv:2011.10098 [Cs]*, Nov. 2020. *arXiv.org*,

Information needed to Achieve Four Degrees of Reproducibility

R1 Description: Only a textual descriptions of the experiment is shared. The text could describe the experimental procedure, the target system and its behaviour, the implementation of the target system for example in form of pseudo code, the data collection procedure, the data, the outcome and the analysis and so on.

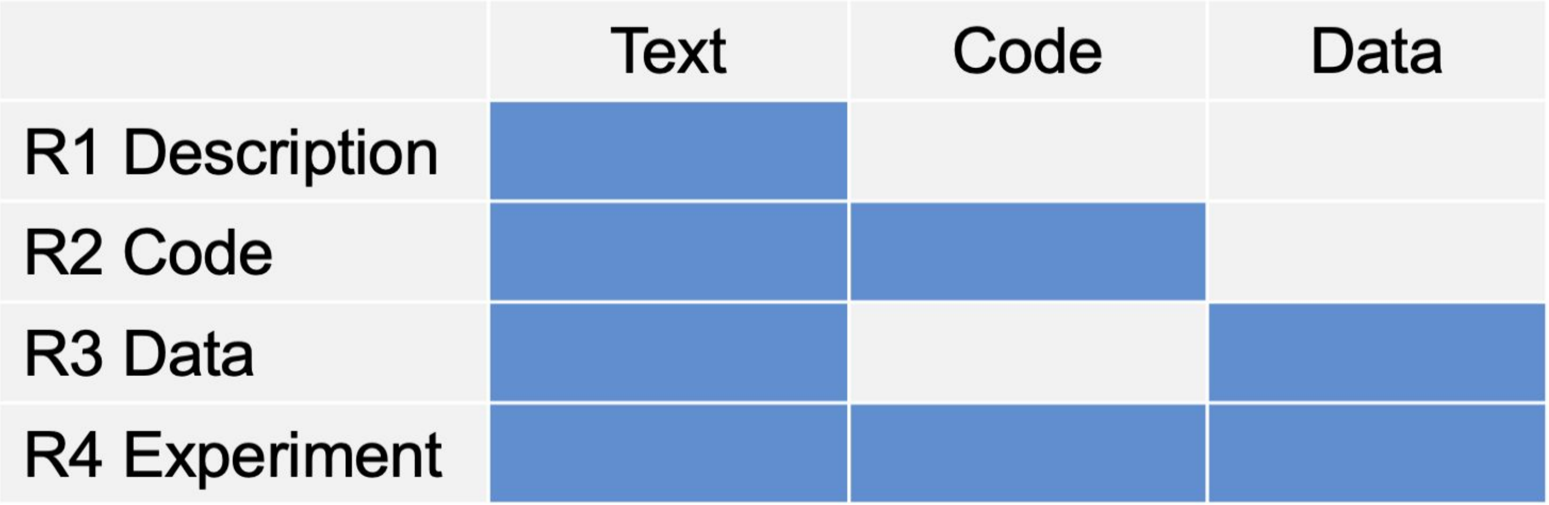
R2 Code: Code is shared in addition to the textual description of the experiment. The code could cover the target system, the workflow, data pre-processing, experiment configurations, visualization and analyses.

R3 Data: Data is shared in addition to the textual description of the experiment. The data could include training, validation and test sets as well as the outcome produced in the experiment.

R4 Experiment: The complete experiment is shared including data and code in addition to the textual description of the experiment.

Gundersen, Odd Erik. "The Fundamental Principles of Reproducibility." *ArXiv:2011.10098 [Cs]*, Nov. 2020. *arXiv.org*,

Four Degrees of Reproducibility



	Text	Code	Data
R1 Description	🟦		
R2 Code	🟦	🟦	
R3 Data	🟦		🟦
R4 Experiment	🟦	🟦	🟦

R4 degree is most likely to be reproducibility. However, lab bias can still make reproducibility difficult.

Gundersen, Odd Erik. “The Fundamental Principles of Reproducibility.” *ArXiv:2011.10098 [Cs]*, Nov. 2020. *arXiv.org*,

Lab Bias

Lab A

Lab B

- **Statistical analysis of the same sample at two labs can differ due to the lab equipment.**
- **Even though the statistical analysis can differ, the conclusions may be the same.**

Source: Miller, J. N., and J. C. Miller. *Statistics and Chemometrics for Analytical Chemistry*. 6. ed. Harlow: Prentice Hall, 2010.

Computer Lab Bias and ML

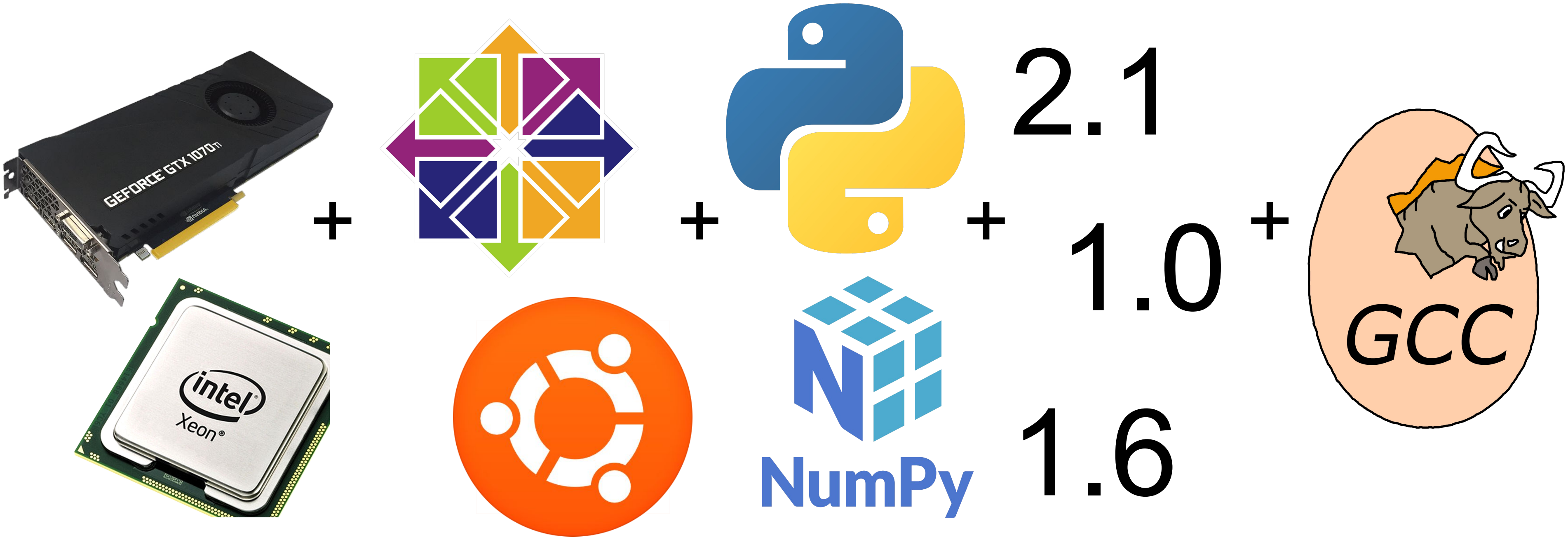
Lab A

Lab B

- Running the same models on different systems produce different results.
- The impact that these biases have on experiment conclusions is understudied.

1. Hong et al., "An Evaluation of the Software System Dependency of a Global Atmospheric Model."

Sources of Variability



Processing Units

Operating Systems

Ancillary Software

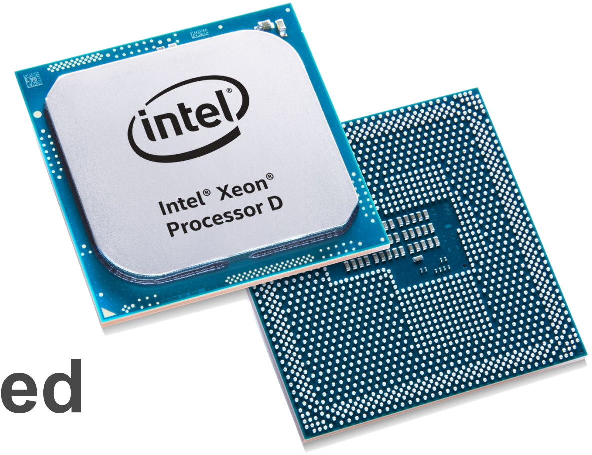
Software Versions

Compiler Options

Processing Units (CPU)

non-associativity of floating point arithmetic

Threading:

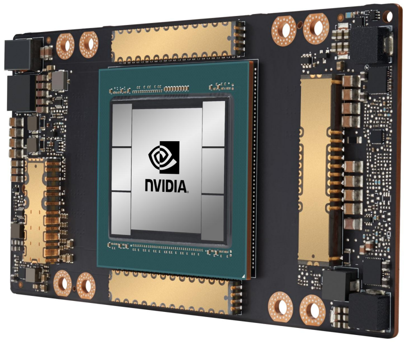


- Given $a + b = e$ & $c + d = f$
- $a + b + c + d = e + f$ is not ensured
- 2 threads: $e + f$
- 4 threads: $a + b + c + d$

Crane, Matt. “Questionable Answers in Question Answering Research: Reproducibility and Variability of Published Results.”

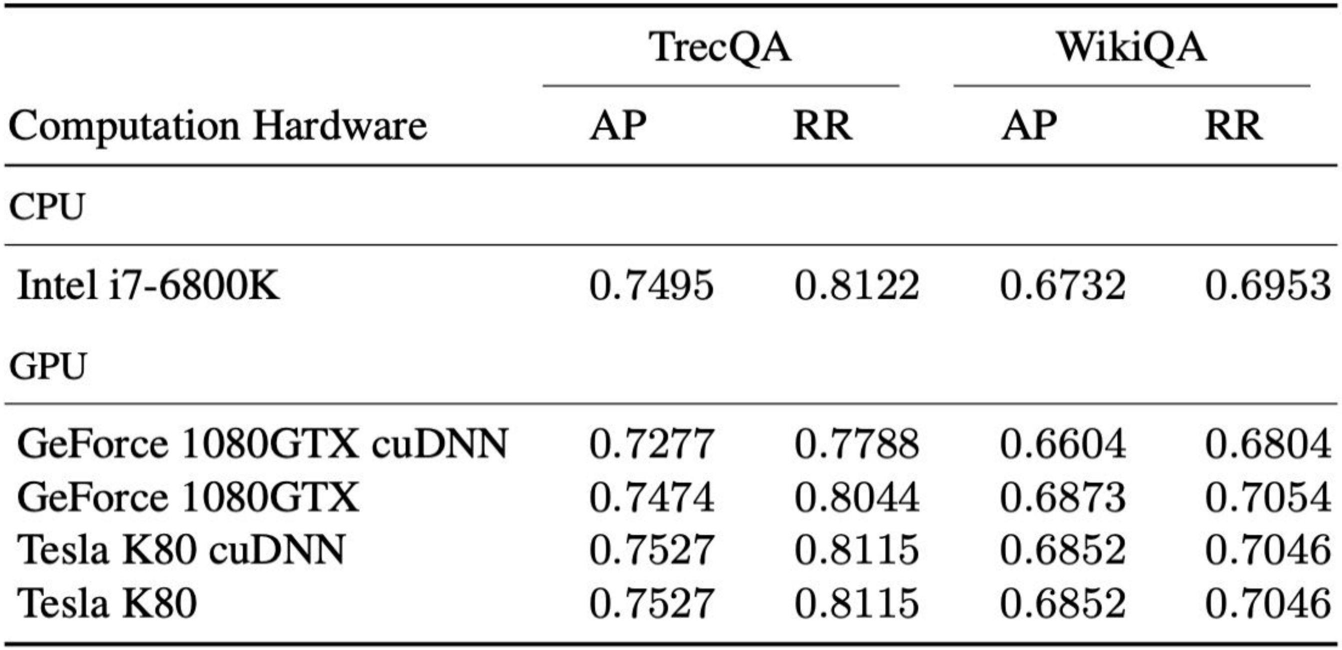
Processing Units (GPU)

- GPUs are in general a greater source of variability than CPUs.
- There are many GPU models and manufacturers are free to deviate from the reference models provided by nVidia or AMD.
- Some factors can not be controlled, for instance, the number of threads that are used by the GPU.



Crane, Matt. "Questionable Answers in Question Answering Research: Reproducibility and Variability of Published Results."

Processing Units (GPU)



	TrecQA		WikiQA	
Computation Hardware	AP	RR	AP	RR
CPU				
Intel i7-6800K	0.7495	0.8122	0.6732	0.6953
GPU				
GeForce 1080GTX cuDNN	0.7277	0.7788	0.6604	0.6804
GeForce 1080GTX	0.7474	0.8044	0.6873	0.7054
Tesla K80 cuDNN	0.7527	0.8115	0.6852	0.7046
Tesla K80	0.7527	0.8115	0.6852	0.7046

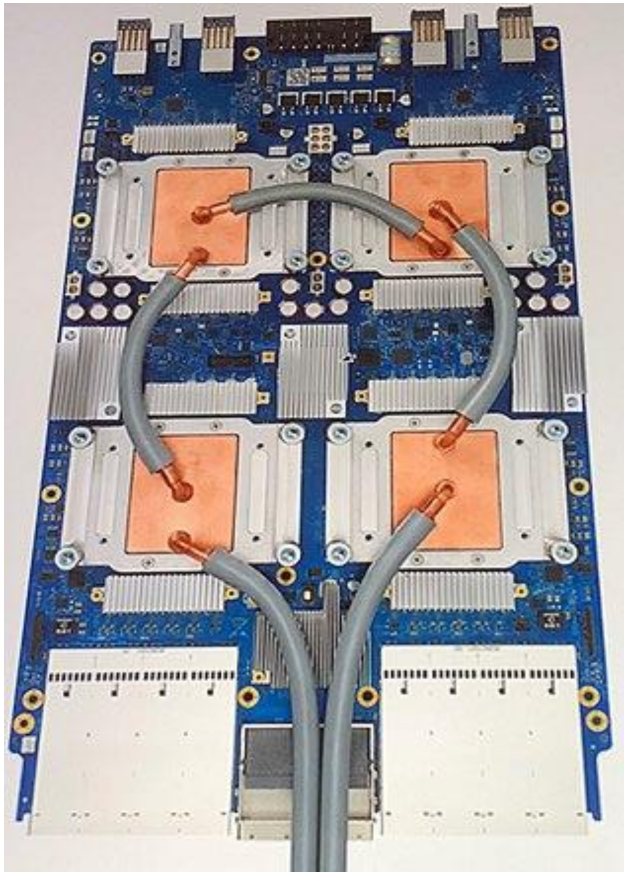
Algorithm for Exploring the Effectiveness of Convolutional Neural Networks for Answer Selection in End-to-End Question Answering is trained and tested using the TrecQA and WikiQA datasets.

average precision (AP) and reciprocal rank (RR)

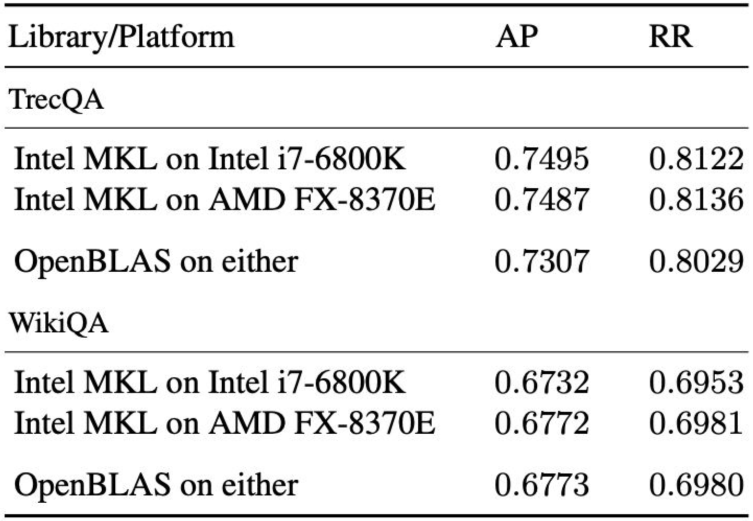
Crane, Matt. "Questionable Answers in Question Answering Research: Reproducibility and Variability of Published Results."

Processing Units (Other)

- **FPGAs & ASICs (TPUs)**
  - The hardware is proprietary
    - Full specs are not open
  - Clones
    - May achieve API compatibility unknown for reproducibility
  - Reproducibility from generation to generation?



Ancillary Software



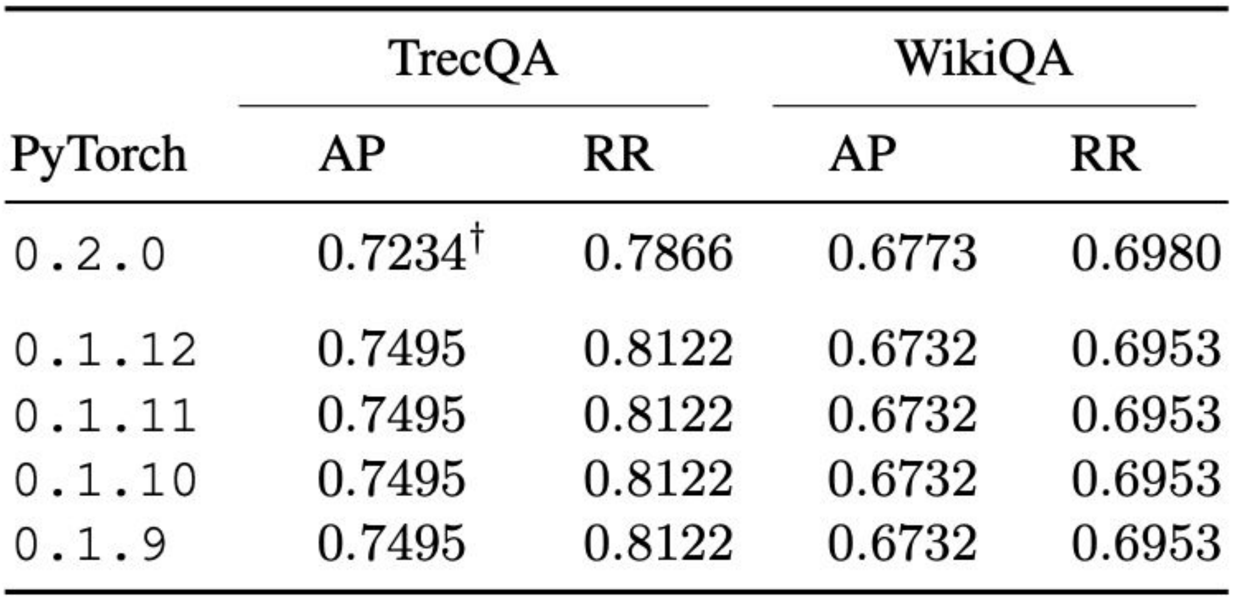
Library/Platform	AP	RR
TrecQA		
Intel MKL on Intel i7-6800K	0.7495	0.8122
Intel MKL on AMD FX-8370E	0.7487	0.8136
OpenBLAS on either	0.7307	0.8029
WikiQA		
Intel MKL on Intel i7-6800K	0.6732	0.6953
Intel MKL on AMD FX-8370E	0.6772	0.6981
OpenBLAS on either	0.6773	0.6980

Effect of changing math library and architecture on model results. Intel MKL on Intel vs AMD vs OpenBLAS on both.

average precision (AP) and reciprocal rank (RR)

Crane, Matt. “Questionable Answers in Question Answering Research: Reproducibility and Variability of Published Results.”

Software Versions



	TrecQA		WikiQA	
PyTorch	AP	RR	AP	RR
0.2.0	0.7234[†]	0.7866	0.6773	0.6980
0.1.12	0.7495	0.8122	0.6732	0.6953
0.1.11	0.7495	0.8122	0.6732	0.6953
0.1.10	0.7495	0.8122	0.6732	0.6953
0.1.9	0.7495	0.8122	0.6732	0.6953

Effect of the version of PyTorch being used on model results.

average precision (AP) and reciprocal rank (RR)

Crane, Matt. “Questionable Answers in Question Answering Research: Reproducibility and Variability of Published Results.”

Compiler Options

New instruction sets are added to x86 processors every generation.

Many of these instruction sets optimize floating point arithmetic.

Differences come from:

- Different approximations to math functions or operations such as division
- The accuracy with which intermediate results are calculated and stored
- Denormalized (very small) results being treated as zero
- The use of special instructions such as fused multiply-add (FMA) instructions

https://techdecoded.intel.io/resources/floating-point-reproducibility-in-intel-software-tools/

Operating Systems

Operating systems are a combination of the Ancillary Software, the Software Versions and the Compiler Options sources of variability.

Randomness in AI

Frequent causes due to the source of variability:

- **GPU:** not guaranteed to generate the bit-wise reproducibility across different GPU versions and architectures.
- **Third-Party Libraries:** libraries used are using other libraries which might in turn use stochastic processes needing a seed to a different random number generator.

Lynnerup, Nicolai A., et al. A Survey on Reproducibility by Evaluating Deep Reinforcement Learning Algorithms on Real-World Robots.

Randomness in Deep Learning

- **Random Initialization of Weights:** The initialization of layer weights of a neural network must be the same from run-to-run in order to expect same results and not similar results.
- **Shuffling of the Datasets:** Datasets are divided into mini-batches and are shuffled in order to avoid the optimization functions getting stuck in local minima.
- **Random Sampling:** Choosing a random subsample from the dataset to train our model with. (too much data)
- **Random Train/Test/Validation Splits:** Using k–fold cross validation where the dataset is stochastically split into two or three sets of data. (too little data)
- **Stochastic Attributes of the Hidden Layers:** One of the most often used techniques for preventing overfitting is dropout which is inherently random during the training process.

Lynnerup, Nicolai A., et al. A Survey on Reproducibility by Evaluating Deep Reinforcement Learning Algorithms on Real-World Robots.

Randomness in Deep Reinforcement Learning

- **Environment:** Especially when dealing with real-world robotic RL, sensor delays, etc. supports the statement that our world is stochastic.
- **Network initialization:** As in DL the initialization of the neural networks' weights are a stochastic process and must thus be controlled for to ensure reproducibility.
- **Minibatch sampling:** Several algorithms within DRL includes sampling randomly from the training data and from replay buffers.

Lynnerup, Nicolai A., et al. A Survey on Reproducibility by Evaluating Deep Reinforcement Learning Algorithms on Real-World Robots.

Quantifying Laboratory Bias

Borrowing from Analytical Chemistry again

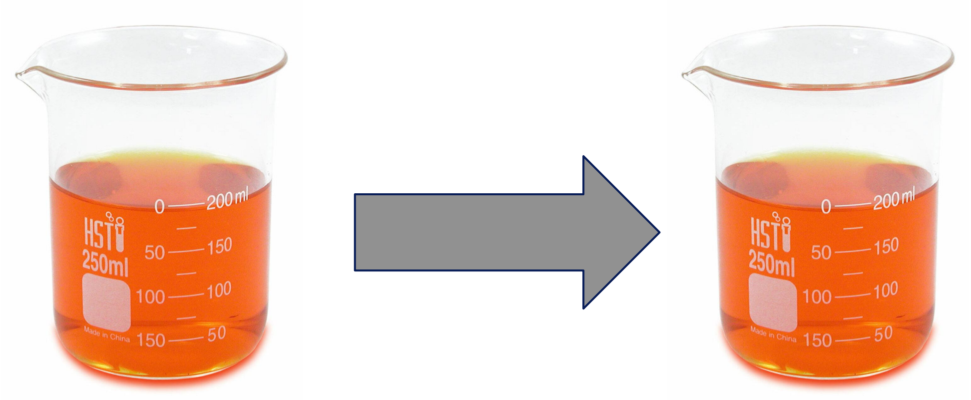
$$S_R = S_r + S_L$$

S_R - Reproducibility standard deviation
S_r - Repeatability standard deviation
S_L - Variance due to inter-laboratory differences, which reflect different degrees of bias in different laboratories

Miller, J. N., and J. C. Miller. *Statistics and Chemometrics for Analytical Chemistry*. 6. ed, Prentice Hall, 2010.

Experiment Reproducibility

Even with laboratory bias and a low degree of documentation, experiments can achieve interpretation reproducibility.



TEAM 1

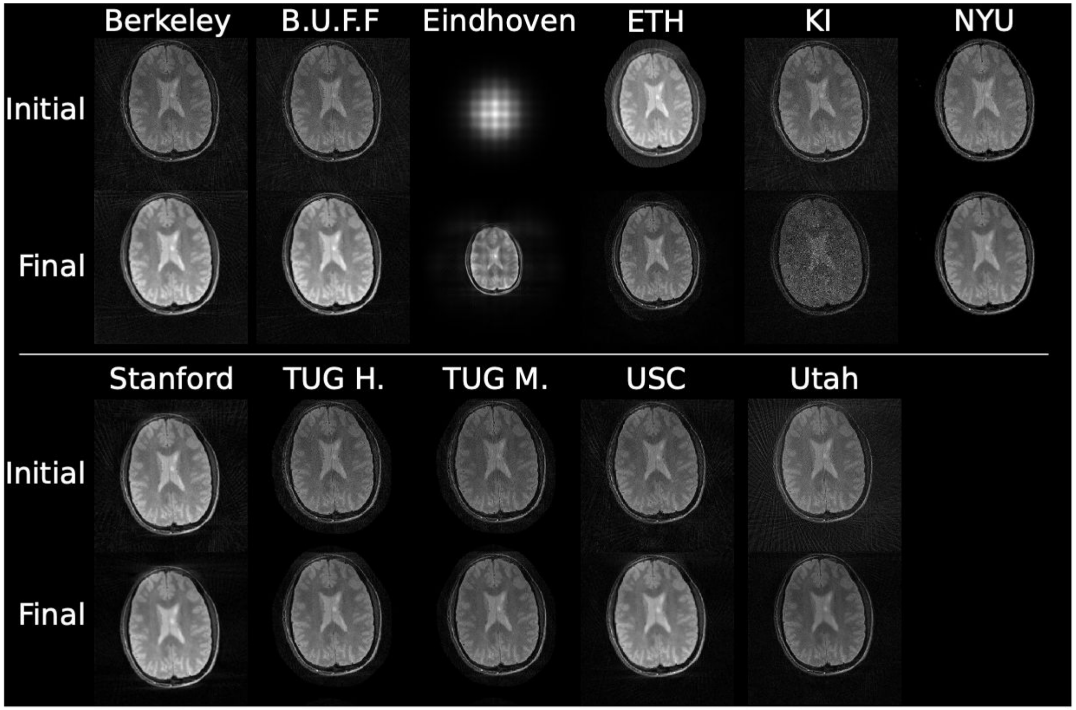
TEAM 2

CG-SENSE ISMRM Reproducibility Challenge

- Recreate the results of "Advances in sensitivity encoding with arbitrary k-space trajectories" by Pruessmann et al. for MR image reconstruction.
- Puressmann et al. only published the algorithm without an example implementation.
- 11 teams from around the world tried to create a implementation of the algorithm.

Maier, Oliver, et al. “CG-SENSE Revisited: Results from the First ISMRM Reproducibility Challenge.” *Magnetic Resonance in Medicine*, vol. 85, no. 4, Apr. 2021, pp. 1821–39. *arXiv.org*, https://doi.org/10.1002/mrm.28569.

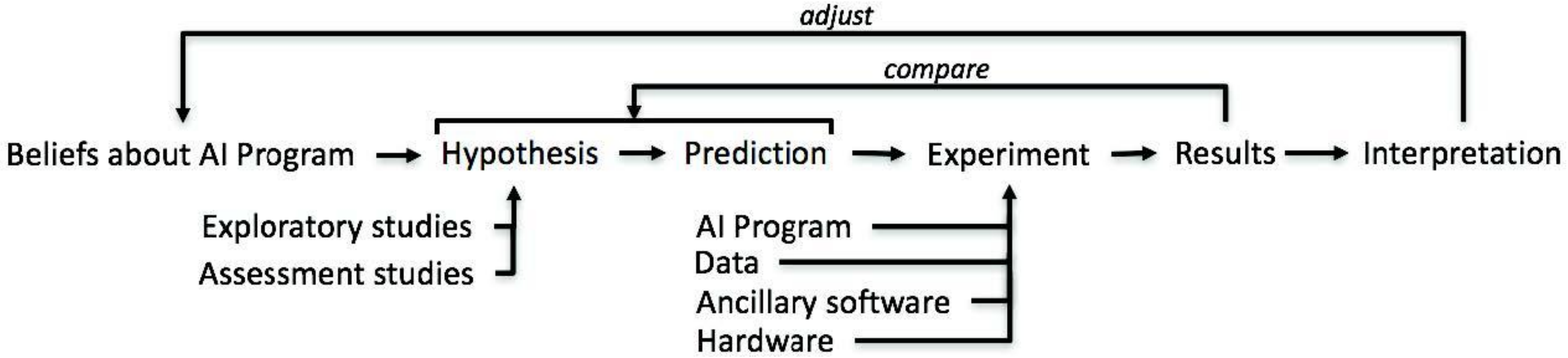
CG-SENSE ISMRM Reproducibility Challenge



The reviewers of the results had concluded that all 11 teams were able to successfully reproduce Pruessmann et al. paper.

Maier, Oliver, et al. “CG-SENSE Revisited: Results from the First ISMRM Reproducibility Challenge.” *Magnetic Resonance in Medicine*, vol. 85, no. 4, Apr. 2021, pp. 1821–39. *arXiv.org*, https://doi.org/10.1002/mrm.28569.

Better Outcomes



adjust

compare

Beliefs about AI Program → Hypothesis → Prediction → Experiment → Results → Interpretation

Exploratory studies

Assessment studies

AI Program

Data

Ancillary software

Hardware

For an experiment to be reproducible, **only the *interpretation* has to be identical, not the results**. Researchers should know the factors that lead to different experiment results when forming their interpretation.

Based on the AI method, does the experiment need:

- Contextual adjustment given the computational environment
- To be conducted in more environments before being fully trusted

Source: Gundersen, Odd Erik, and Sigbjorn Kjensmo. “State of the Art: Reproducibility in Artificial Intelligence,” n.d., 8.

Contact Information

You can contact me if you have any questions, thoughts or suggestions:

Kevin Coakley - kcoakley@sdsc.edu

Thank you!